On Bridging the Gap between Mean Field and Finite Width Deep Random Multilayer Perceptron with Batch Normalization

Amir Joudaki [1] **Hadi Daneshmand** [2 3 4] **Francis Bach** [5]

Abstract

Mean-field theory is widely used in theoretical studies of neural networks. In this paper, we analyze the role of depth in the concentration of mean-field predictions for Gram matrices of hidden representations in deep multilayer perceptron (MLP) with batch normalization (BN) at initialization. It is postulated that the mean-field predictions suffer from layer-wise errors that amplify with depth. We demonstrate that BN avoids this error amplification with depth. When the chain of hidden representations is rapidly mixing, we establish a concentration bound for a mean-field model of Gram matrices. To our knowledge, this is the first concentration bound that does not become vacuous with depth for standard MLPs with a finite width.

1. Introduction

There is a growing demand for a theoretical understanding of neural networks to improve their safety, robustness, computational and statistical effectiveness. Originating in statistical mechanics for investigating complex systems with interacting particles, this theory has been repurposed in recent years for exploring neural network dynamics under the regime of infinite width. By going beyond the microscopic changes of individual neurons, mean field analysis has revealed the collective neuronal behaviors that emerge at initialization (Pennington et al., 2018; Yang et al., 2019; Pennington & Worah, 2017), throughout training (Jacot et al., 2018; Chizat & Bach, 2018; Lee et al., 2019), and after training (Chizat & Bach, 2020; Ba et al., 2019).

[1]Department of Computer Science, ETH Zurich [2]Laboratory for Information and Decision Systems (LIDS), MIT [3]Foundation of Data Science Institute (FODSI) [4]Hariri Institute for Computing and Computational Science and Engineering, Boston University [5]INRIA-ENS-PSL Paris. Correspondence to: Amir Joudaki <amir.judaki@gmail.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

In this paper, we delve into the role of mean field theory at initialization when network weights are allocated randomly. Pioneering works by Glorot & Bengio (2010) and Saxe et al. (2013) underscored the impact of initialization on training, paving the way for mean field theory to uncover a wealth of insights. Examples include identifying connections between wide neural networks and Gaussian processes (Neal, 1995; de G. Matthews et al., 2018; Jacot et al., 2018), examining the concentration of singular values of input-output Jacobians (Pennington et al., 2018; Feng et al., 2022), and designing activation functions (Klambauer et al., 2017; Ramachandran et al., 2017; Li et al., 2022). Remarkably, Xiao et al. (2018) introduced an initialization scheme capable of training convolutional networks comprising 10000 layers.

A common thread among these studies is the dynamics of inner products between hidden representations, encoded in their Gram matrix. Mean field theory models these dynamics via a difference equation derived from the infinite-width limit of the network. However, mean field analysis is inherently prone to approximation errors when dealing with networks of finite width. As de G. Matthews et al. (2018) observed, this error grows with depth, ultimately leading to vacuous error bounds in the infinite depth limit. To tackle this issue, they propose to increase the network width proportional to depth. This idea is echoed in other studies which propose maintaining a constant ratio between depth and width (Hanin & Nica, 2019; Li et al., 2021), a regime in which Hanin (2022) confirmed a constant concentration bound for mean field estimates.

Can we achieve a bounded mean field error when the width is finite? We answer this question affirmatively for MLPs that are endowed with batch normalization. In particular, we show that under some technical assumption on the underlying dynamics (as formally expressed in Theorem 1), the mean field estimation error for Gram matrices remains bounded at infinite depth. Specifically, we demonstrate that this error is limited by $\text{width}^{-1/2}$ with high probability. This contrasts with the vacuous concentration bounds at infinite depth observed in the absence of normalization (Li et al., 2022). Our results highlight the importance of existing mean field analyses of batch normalization by Yang et al. (2019), and demonstrate their high accuracy in the finite

width scenarios that are relevant for practical applications.[1]

2. Related works

Numerous studies (Saxe et al., 2013; Feng et al., 2022; Yang et al., 2019) have provided valuable insights into training neural networks by studying input-output Jacobians of neural networks with and without normalization at initialization. For example, Feng et al. (2022) have shown that the rank of the input-output Jacobian of neural networks without normalization at initialization diminishes exponentially with depth, while Yang et al. (2019) have shown that batch normalization avoids this exponential diminishing.

The spectrum of Jacobians is intimately related to the spectra of Gram matrices. A Gram matrix (G-matrix) contains inner products of samples within a batch (equation 2). Thus, a degenerate G-matrix for the penultimate layer implies that the outputs are insensitive to the inputs (Feng et al., 2022; Li et al., 2022). Rank collapse in the last hidden layer occurs in various neural network architectures, including MLPs (Saxe et al., 2013), convolutional networks (Daneshmand et al., 2020), and transformers (Dong et al., 2021), and leads to ill-conditioning of the input-output Jacobian, which slows training (Daneshmand et al., 2021; Pennington et al., 2018; Yang et al., 2019). Saxe et al. (2013) have shown that avoiding rank collapse can accelerate the training of deep linear networks, making it a focus of theoretical and experimental research (Pennington et al., 2018; Daneshmand et al., 2020; 2021).

A recent line of research (Daneshmand et al., 2020) postulates that batch normalization can enhance the training of deep neural nets by avoiding the rank collapse. This claim has been supported by empirical evidence (Yang et al., 2019; Daneshmand et al., 2020), as well as theoretical studies for neural networks with infinite widths (Yang et al., 2019) and linear activation (Daneshmand et al., 2021). It has been shown that batch normalization prevents degenerate representations at initialization (Daneshmand et al., 2020), and orthogonalizes representations (Daneshmand et al., 2021). However, these results are limited to linear activations. The present study extends these findings to neural networks with finite widths and non-linear activations, under an assumption from Markov chain theory.

3. Problem settings and background

Notation and terminology. I_n denotes the identity matrix of size $n \times n$ and 1_n denotes the all ones vector in $\mathbb{R}^n$. $\otimes$ refers to Kronecker product. μ_X refers to the probability measure of the random variable X. We use $f \lesssim g, g \gtrsim f$ and $f = O(g)$ to denote the existence of an absolute constant c such that $f \leq c\ g$. $\|v\|$ for a vector v denotes the L^2 norm. $\|C\|$ for matrix C denotes the L^2 operator norm $\|C\| = \sup_{x\in\mathbb{R}^n} \|Cx\|/\|x\|$, $\|C\|_F$ denotes the Frobenius norm. We use $\kappa(C)$ to denote the ratio of largest to smallest eigenvalue. Both $h_{r\cdot}$ and $\mathrm{row}_r(h)$ denote row-vector representation of the r-th row of h. Finally, $X \sim \mathcal{N}(\mu, \sigma^2)^{n\times m}$ denotes $X \in \mathbb{R}^{n\times m}$ is a Gaussian matrix whose elements are drawn i.i.d. from $\mathcal{N}(\mu, \sigma^2)$.

Setup. Let $h_\ell \in \mathbb{R}^{d\times n}$ denote the hidden representation at layer ℓ, where n corresponds to the size of the mini-batch, and d denotes the width of the network that is kept constant across all layers. The sequence $\{h_\ell\}$ is a Markov chain as

$$h_{\ell+1} := W_\ell \sigma \circ \phi(h_\ell), \qquad W_\ell \sim \mathcal{N}(0, 1/d)^{d\times d}, \tag{1}$$

where $h_0 \in \mathbb{R}^{d\times n}$ is the input batch, σ is the element-wise activation function, and ϕ is the batch normalization (Ioffe & Szegedy, 2015), which applies row-wise centering and scaling by standard deviation:

$$\phi(x) = \frac{x - \mathrm{mean}(x)}{\sqrt{\mathrm{Var}(x)}}, \quad \forall r : \mathrm{row}_r(\phi(h)) = \phi(\mathrm{row}_r(h)).$$

4. Mean-field models and fixed-point analyses

4.1. Mean-field Gram Dynamics

The Gram matrix G_ℓ is defined as the matrix of inner products of hidden representations at layer ℓ as seen in the equation below:

$$G_\ell := \frac{1}{d}(\sigma \circ \phi(h_\ell))(\sigma \circ \phi(h_\ell))^\top. \tag{2}$$

Understanding the dynamics of G_ℓ is a significant challenge in deep learning theory, and has been the subject of several studies (Yang et al., 2019; Pennington et al., 2018; Pennington & Worah, 2017). Due to the randomness of weights, determining the trajectory of this random process proves to be arduous. By tending width d to infinity, i.e., the mean-field regime, we can approximate these stochastic dynamics with a deterministic dynamics as below:

$$\overline{G}_{\ell+1} = \mathbb{E}_{h\sim\mathcal{N}(0,\overline{G}_\ell)} \left[\sigma \left(\frac{\sqrt{n} M h}{\|Mh\|} \right)^{\otimes 2} \right], \tag{3}$$

Where $\overline{G}_0 = G_0$ serves as the input G-matrix and $M = I_n - \frac{1}{n} 1_n^{\otimes 2}$ applies mean reduction on the preactivations. The mean-field approach in this context assists in elucidating the analysis of Gram matrices.

4.2. Fixed point analysis for infinitely deep and wide Networks

The fixed points of the mean field dynamics, as expressed in equation 3 may help elucidate the properties of $\overline{G}_\ell$ as

[1] Codes available at https://github.com/ajoudaki/mean-field-normalization.

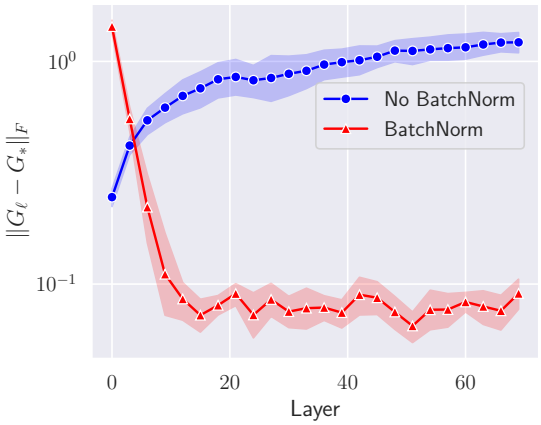

Figure 1. Mean field error amplification with(out) batch-normalization. The horizontal axis represents the number of layers ℓ (linear), while the vertical axis (log-scale) shows $\|G_\ell - G_*\|_F$, for networks with $n = 5, d = 1000$. The traces show mean and shades indicate 90% confidence intervals over 10 independent simulations.

$\ell \to \infty$. Yang et al. (2019) provide a comprehensive characterization of these fixed-points, denoted by G_*, for neural networks with batch normalization. For networks with linear activations, Yang et al. (2019) establish a global convergence to these well-conditioned fixed-points. While they empirically observe convergence to these well-conditioned fixed-point for networks with non-linear activations, that is not established theoretically. In other words, it is challenging to describe the properties of $\overline{G}_\ell$ for finite width and depth, and it is unclear how the fixed-point Gram matrix G_* can inform us about G_ℓ.

4.3. An observation

Through an empirical observation we can demonstrate that G_* may not always provide an accurate estimate for G_ℓ. We observe that for a network without batch normalization and linear activations (when $\sigma \circ \phi = identity$ in equation 1), the Frobenius distance between G_ℓ and G_* increases with ℓ. In contrast, G_ℓ converges to a neighborhood of G_* when the network includes batch normalization layers. These observations suggest that the mean field estimate G_* from Yang et al. (2019) accurately represents G_ℓ when batch normalization is present.

4.4. The challenge of depth for mean-field theory

Mean field analysis suffers from a systematic estimation error that increases with depth. Assuming that $\overline{G}_0 = G_0$, then an error of $O(d^{-1/2})$ is observed between $\overline{G}_1$ and G_1 due to the concentration of empirical covariance. Consequently, the mean-field dynamics in equation 3 incur an error of $O(d^{-1/2})$ at each layer (Li et al., 2022). As depth ℓ grows, these errors are amplified, and the bounds on $\|G_\ell - \overline{G}_\ell\|_F$ become vacuous, thus raising questions about the practical applicability of these fixed-point analyses when width is finite.

Several studies strive to refine the mean-field model to enhance its predictive accuracy (de G. Matthews et al., 2018; Hanin & Nica, 2019; Li et al., 2022). Li et al. (2022) propose using a stochastic differential equation to model the layer-wise $O(d^{-1/2})$ estimation error for mean-field Gram dynamics. This approach allows for accurate predictions of Gram dynamics for MLPs with activation functions but only in the infinite-width-and-depth regime. Our observations, however, suggest that for networks with batch normalization, the deterministic model of Gram matrices provides a surprisingly accurate estimate.

Daneshmand et al. (2021) established this observation for multilayer perceptrons (MLPs) with batch normalization (BN) and linear activations, subject to specific conditions. They demonstrated that as ℓ increases, batch normalization progressively aligns the Gram matrices G_ℓ with the identity matrix, which coincides with G_* for such networks (Yang et al., 2019). Yang et al. (2019) further proved a concentration bound for $\|G_\ell - G_*\|_F$ in networks with batch normalization and linear activations. However, both these findings are limited to linear activations. Our objective is to extend these results to networks incorporating non-linear activations.

5. Concentration bounds for Mean-field Predictions with Batch Normalization

5.1. Geometric ergodic assumption

The chain of hidden representations obeys a non-linear stochastic recurrence. Despite this non-linearity, the distribution associated with the representation obeys a linear fixed-point iteration determined by the Markov kernel K associated with the chain h_ℓ. The distribution of h_ℓ, denoted by μ_ℓ, obeys

$$\mu_{\ell+1} = T(\mu_\ell), \quad T(\mu) := \int K(x, y) d\mu(y). \tag{4}$$

The fixed-points of the above equation are invariant distributions of the chain, which we denote by μ_*. Recall that the total variation for distributions over $d \times n$ matrices can be defined as $\|\mu_X - \mu_Y\|_{tv} := \sup_{A \subseteq \mathbb{R}^{d \times n}} |\mu_X(A) - \mu_Y(A)|$. Notably, the above recurrence is non-expansive in total variation, hence $\|\mu_\ell - \mu_*\|_{tv} \leq \|\mu_{\ell-1} - \mu_*\|_{tv}$ holds for all ℓ. However, we assume the chain obeys a strong property ensuring the convergence to a unique invariant distribution.

Assumption 1 (Geometric ergodicity)**.** *We assume the chain of hidden representations admits a unique invariant distribution. Furthermore, there is constant α ($\alpha > 0$) such*

that

$$\|\mu_\ell - \mu_*\|_{tv} \leq (1-\alpha)^\ell \|\mu_0 - \mu_*\|_{tv},$$

holds almost surely for all h_0.

The geometric ergodic property is established for various Markov chains, such as the Gibbs sampler, state-space models (Eberle, 2009), hierarchical Poisson models (Rosenthal, 1995), and Markov chain Monte Carlo samplers (Jones, Galin L. and Hobert, James P., 2001). Doeblin (1938) provides weak conditions that ensure geometric ergodicity. Doeblin's condition holds when the Markov chain can explore the entire state space (Eberle, 2009). This condition may hold under weak assumption on the input matrix for the chain of hidden representations. In particular, when h_ℓ has full rank, the Gaussian product $W_\ell h_\ell$ may explore the entire $\mathbb{R}^{d\times n}$.

5.2. Main result

The next theorem proves fixed-point G_* provides an estimate for Gram matrices of sufficiently deep neural networks with a finite width.

Theorem 1 (BN-MLP Concentration)**.** *Assume the Markov chain of representations* $\{h_\ell\}$ *obeys Assumption 1 with* $\alpha > 0$, *and has non-degenerate fixed-point* G_*. *If the activation* σ *is uniformly bounded* $|\sigma(x)| = O(|x|)$, *then Gram matrix deviation* $\|G_* - G_\ell\|_F$ *is bounded by*

$$\kappa(G_*) O\left((1-\alpha)^{\frac{\ell}{2}} + \frac{n}{\sqrt{d}}\alpha^{-\frac{1}{2}} \ln^{\frac{1}{2}}(\frac{d}{n})\right), \tag{5}$$

with high probability in d *and* ℓ.

Theorem 1 quantifies the accuracy of our mean-field predictions in terms of batch size, width, depth, and conditioning of G_*. Notably, almost all commonly used activations, e.g., ReLU and hyperbolic tangent, satisfy the uniform bounded condition $|\sigma(x)| = O(|x|)$. Under Assumption 1, this theorem proves the fixed-point Gram matrix G_* accurately estimates G_ℓ for a sufficiently large ℓ. According to this theorem, $\|G_\ell - G_*\|_F$ decays with depth at an exponential rate. Thus, approximately after a logarithmic number of layers $\ell \approx \log(\text{width}/\text{batch-size})$, the term $O(n/\sqrt{d})$ dominates the distance.

Remarkably, this is a considerable improvement compared to the concentration bounds for neural networks without batch normalization that become vacuous as the depth increases (Hanin & Nica, 2019; Hanin, 2022). The established bound in the last theorem holds jointly for all ℓ in that we do not need to apply union bound.

Let us remark that if the fixed-point Gram matrix is degenerate, i.e., if $\kappa(G_*)$ is unbounded, the bound of the Theorem becomes vacuous. Therefore, Theorem 1 reinforces the necessity for a well-conditioned fixed-point for the mean-field errors to remain within bounds. As long as the fixed-point Gram is well-conditioned, the Gram matrices G_ℓ's stay within an $O(\text{batch}/\text{width}^{1/2})$ proximity with constant probability.

When contrasting Theorem 1 with the activation shaping approach by Li et al. (2022), we observe that while activation shaping necessitates solving a stochastic differential equation to track the dynamics of the Gram matrix, BN-MLP relies solely on the mean-field prediction G_*, which can be computed in closed-form Yang et al. (2019).

Proof Sketch of Theorem 1. We first construct an approximate invariant distribution, associated with T as defined in equation 4. For the construction of such distribution, we utilize the mean-field Gram matrix to form an input $\hat{h} \in \mathbb{R}^{d\times n}$, with rows drawn i.i.d. from $\text{row}_r(\hat{h}) \sim \mathcal{N}(0, G_*)$. The next lemma proves that the law of $\hat{h}$, denoted by $\hat{\mu}$, does not significantly change under T.

Lemma 2. *Assuming uniformly bounded activation* $|\sigma(x)| = O(|x|)$, *we have*

$$\|T(\hat{\mu}) - \hat{\mu}\|_{tv} \lesssim \|G_*^{-1}\| \frac{n^2}{d} \ln(d/n). \tag{6}$$

The proof of the last lemma is based on the fixed-point property of G_* (see Appendix for the detailed proof). Using the last lemma together with Assumption 1, we prove that $\hat{\mu}$ is in a tv-ball around the invariant distribution μ_*. Under this assumption, we have

$$\begin{aligned} \|T(\hat{\mu}) - T(\mu_*)\|_{tv} &= \|T(\hat{\mu}) - \mu_*\|_{tv} \\ &\leq (1-\alpha)\|\hat{\mu} - \mu_*\|_{tv}, \end{aligned} \tag{7}$$

where we used the invariant property of μ_* in the above equation. Using triangular inequality, we get

$$\begin{aligned} \|T(\hat{\mu}) - \hat{\mu}\|_{tv} &= \|T(\hat{\mu}) - \mu_* + \mu_* - \hat{\mu}\|_{tv} \\ &\geq \|\mu_* - \hat{\mu}\|_{tv} - \|T(\hat{\mu}) - \mu_*\|_{tv} \\ &\geq \alpha\|\hat{\mu} - \mu_*\|_{tv}. \end{aligned} \tag{8}$$

Plugging the bound from the last lemma into the above inequality concludes $\hat{\mu}$ lies within a radius $n^2\|G_*^{-1}\|/d\alpha$ of μ_*. This concludes the proof: Since the chain is geometric ergodic, the distribution μ_ℓ converges to an tv-ball around $\hat{\mu}$ at an exponential rate. This allows us to characterize the moments of μ_ℓ using those of $\hat{\mu}$.

5.3. Validation of main theoretical results

Our principal finding suggests a link between the Gram matrices of hidden representations with independent weights. Assuming $\kappa(G_*) = O(1)$ this is captured by the relation:

$$\|G_\ell - G_*\|_F = O\left((1-\alpha)^{\ell/2} + \frac{n}{\sqrt{d}}\right). \tag{9}$$

We test this relationship by numerically estimating G_* by tending d and ℓ to sufficiently large values. We then plot the left-hand side of the above equation versus depth, width, and batch size in the following figures. These plots illustrate how the difference in Gram matrices changes with respect to depth, width, and batch size. This supports our theoretical results and showcases their potential implications for practical settings.

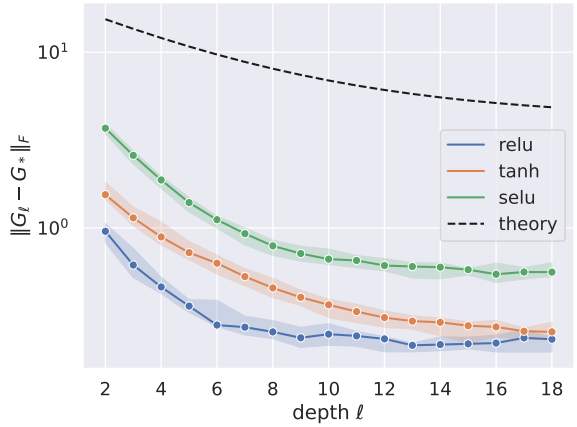

Figure 2. $\|G_\ell - G_*\|_F$ vs. depth, $\ell = 1, 2, ..., 20$, with a fixed width of $d = 1000$ and a batch size of $n = 10$. The dashed line shows the theoretical upper bound of Theorem 1.

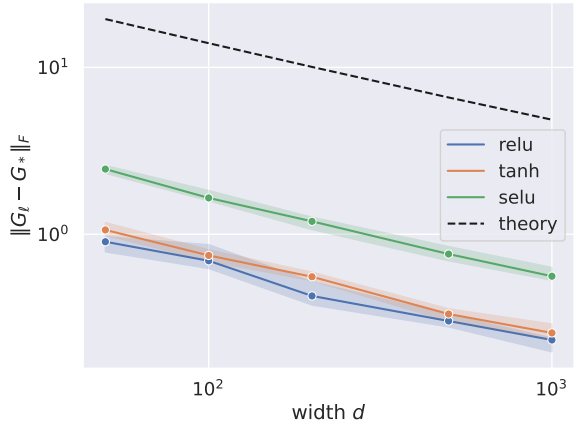

Figure 3. $\|G_\ell - G_*\|_F$ vs. width, $d = 50, 100, 200, 500, 1000$, with a fixed depth of $\ell = 20$ and a batch size of $n = 10$. The second term $O(n/\sqrt{d})$ is always dominant, as demonstrated in the following log-log plot.

6. Applications

The spectral analysis of gram matrices plays a central role in numerous theoretical and practical machine learning studies. For instance, these matrices are used to design activations, contributing to improved conditioning (Klambauer et al., 2017), and to create novel initialization schemes for training convolutional networks with 10000 layers (Xiao et al., 2018). One line of research links the enhanced performance of neural networks incorporating batch normalization to the well-conditioning of Gram matrices G_ℓ (Yang et al., 2019; Daneshmand et al., 2020; 2021). Since the existing literature often uses mean-field approximations, we can leverage Theorem 1 to evaluate accuracy of these approximations for finite width and depth settings.

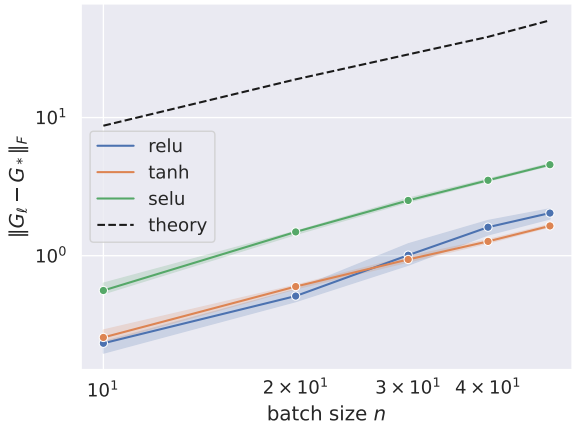

Figure 4. $\|G_\ell - G_*\|_F$ vs. batch size, with a fixed width of $d = 1000$ and a depth of $\ell = 20$, and varying batch sizes of $n = 10, 20, 30, 40, 50$. Dashed line shows upper bound given in Theorem 1.

6.1. Well-conditioning with batch normalization

Empirical studies suggest that the conditioning of Gram matrices, G_ℓ, has a substantial impact on the training of deep neural networks (Xiao et al., 2018; Pennington et al., 2018; Li et al., 2022; Daneshmand et al., 2020). Experimental evidence suggests that batch normalization can ensure the good conditioning of G_ℓ (Yang et al., 2019; Daneshmand et al., 2021), thereby enhancing the training of deep neural networks.

In a seminal study, Yang et al. (2019) study the fixed points of the mean-field equation of an MLP with batch normalization. In particular, they demonstrate that one fixed point G_* follows the following form:

$$G_* = b^* \left((1 - c^*) I_n + c^* \mathbf{1}_{n \times n} \right), \tag{10}$$

where c^* and b^* are constants determined by the activation function. Given the distinctive construction for G_*, we can deduce the structure of its eigenvalues, with the largest eigenvalue being $\lambda_1^* = (1 + (n-1)c^*)b^*$, and all others being equal to $\lambda_2^* = \cdots = \lambda_n^* = b^*(1 - c^*)$.

While the mean field analysis discussed above holds for infinitely wide and deep neural networks, it is possible to utilize Theorem 1 to link the spectrum of G_* with the spectra of Gram matrices for networks of finite width. Using the

Hoffman-Wielandt inequality (Hoffman & Wielandt, 2003), we can calculate a bound on the deviation of the spectrum of G_ℓ from G_*, using the bound on their Frobenius distance.

Corollary 3 (Spectral concentration)**.** *In the same settings as Theorem 1, let λ_i and λ_i^* denote eigenvalues of G_ℓ and G_* respectively in a descending order. Assuming that $\kappa(G_*) = O(1)$, the deviation of their spectra $\sqrt{\Sigma_{i=1}^n (\lambda_i - \lambda_i^*)^2}$ is bounded by*

$$O\left((1-\alpha)^{\frac{\ell}{2}} + \frac{n}{\sqrt{d}}\alpha^{-\frac{1}{2}} \ln^{\frac{1}{2}}(\frac{d}{n})\right), \tag{11}$$

with high probability in d and ℓ.

Substituting the spectrum of G_* characterized by Yang et al. (2019) into the above concentration, we can estimate the spectra of Gram matrices G_ℓ, encapsulated in the following proposition.

Proposition 4. *In the same setting as Theorem 1, assuming G_* complies with equation 10, then for a sufficiently deep layer ℓ, $n - O(1)$ eigenvalues of G_ℓ are within $O(\sqrt{n/d})$ range of $b^*(1-c^*)$ with high probability in d.*

The above proposition provides a characterization for the "bulk" of eigenvalues of G_ℓ, by postulating that majority of eigenvalues of G_ℓ are concentrated around some absolute constant, up to $\sqrt{n/d}$ range. Interestingly, this bears resemblance to the Marchenko-Pastur (Pastur & Martchenko, 1967) law on the eigenvalue distribution of Wishart matrices of comparable size. We observe empirically that the singular values of h_ℓ, which are square root of eigenvalues of G_ℓ, accurately follow the Marchenko-Pastur distribution with $\gamma = n/d$, as depicted in Figure 5. Our empirical evaluations show that this distribution accurately predicts singular values of hidden representations for commonly used activation functions with various widths (see Appendix for empirical evidence).

It is worth noting that only the $O(1)$ singular values are influenced by the activation function, while the remaining $n - O(1)$ exhibit a universal behavior. For example in the case of $\sigma = \mathrm{relu}$, a single large eigenvalue is associated with the $\mathbf{1}_n$ direction, owing to the non-negativity of relu outputs.

6.2. Influence of Gram matrix conditioning on training

Having explored the influence of batch normalization on the spectra of the Gram matrices, we now turn our attention to its effects during training. It has been hypothesized that batch normalization facilitates the training of neural networks at the initialization stage by ensuring the deep representations are biased towards a uniform prior on class probabilities (Daneshmand et al., 2021). In contrast, it has been reported poor Gram matrix conditioning in standard

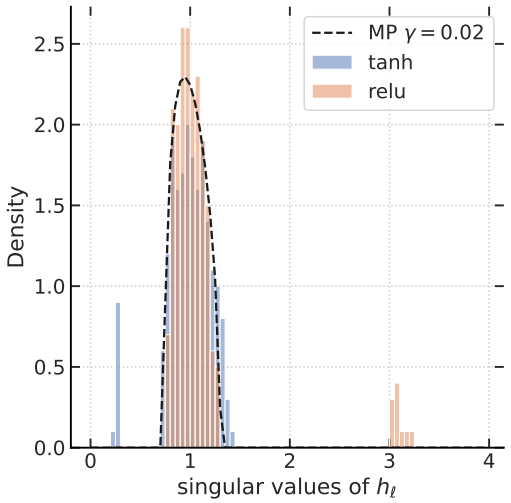

Figure 5. BN-MLP with $n = 20, d = 1000$, $\ell = 20$: histogram shows the empirical distribution of singular values of h_ℓ, for $\sigma = \mathrm{relu}$ and $\sigma = \tanh$. The black curve marks the Marchenko-Pastur distribution with $\gamma = n/d = 0.02$. The singular values are normalized by their medians in this plot to be aligned at 1.0

MLP leads to the gradual alignment of deep hidden layers, thereby resulting in highly similar logits across different samples in the batch (Daneshmand et al., 2020; 2021). Batch normalization effectively resolves this issue, ensuring a more efficient learning process for the network. While our theoretical studies are limited to initialization, we can empirically explore conditioning of Gram matrix during training.

We examined a standard MLP setup consisting of 10 layers ($L = 10$) and a width of 1000 ($d = 1000$). We trained this network on mini-batches of size 128 ($n = 128$) using the CIFAR100 dataset for 50 epochs, using SGD with a learning rate of 0.001. This process was carried out on MLP configurations both with and without batch normalization.

We present the distributions of log-eigenvalues for the penultimate Gram matrix, represented by $\log(\lambda_i(G_L))$, during training in Figure 6. It is noteworthy that the eigenvalues of the penultimate Gram matrix for the MLP with batch normalization are more concentrated around their mean than their counterparts in the MLP without batch normalization. This suggests a collapse in class representations in the absence of normalization.

To further investigate class representations, we calculated the frequency of each class in the predictions at different training stages. We quantified the entropy of the predicted class probabilities, computed as $\sum_{i=1}^C p_i \log_2(p_i)$. In this equation, $p_c := \frac{1}{N}\#\{i \leq N : \tilde{y}_i = c\}$ designates the proportions of predictions for class c, and $\tilde{y}_i$ represents the prediction for sample i. Observe that a uniform distribution $p_1 = \cdots = p_C = 1/C$, leads to the highest entropy

$O(\log_2(C))$.

As illustrated in Figure 7, the MLP with batch normalization closely approximates this uniform prior at initialization. In contrast, the MLP without normalization exhibits a significantly lower initial class entropy. For a balanced dataset like CIFAR100, an optimal model should have a class entropy of approximately $\log_2(100) \approx 6.64$, reflecting a uniform distribution over classes. Hence, batch normalization biases the initial predictions towards a uniform distribution on the labels. The observed discrepancy in Figure 7 may thus be related to the accelerated convergence of training loss as depicted in Figure 8.

The empirical evidence presented here suggests that while Theorem 1 was proven for initialization, Gram matrices of MLP with batch normalization remain well-conditioned during the entire training process.

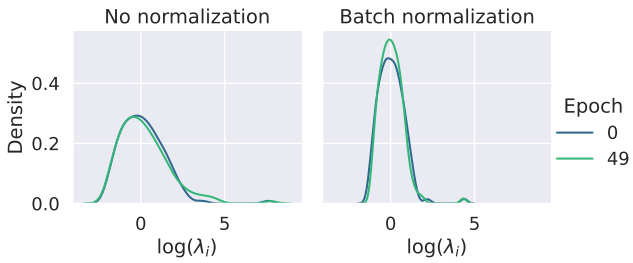

Figure 6. Distribution of log-eigenvalues of penultimate Gram matrix, $\log(\lambda_i(G_L))$ at initialization (epoch 0) and at the end of training (epoch 49).

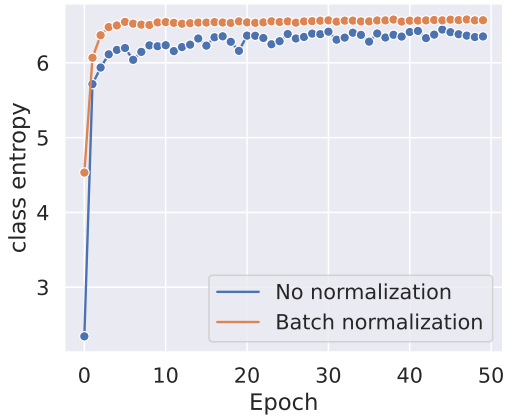

Figure 7. Predicted class entropy of MLP with (orange) and without (blue) batch normalization.

7. Limitations and Future Directions

In this paper, we presented a theoretical framework that bridges the gap between the mean-field theory of neural networks with finite and infinite widths, with a focus on batch normalization at initialization. Many questions that were out of the scope for this study, suggesting directions for new lines of inquiry.

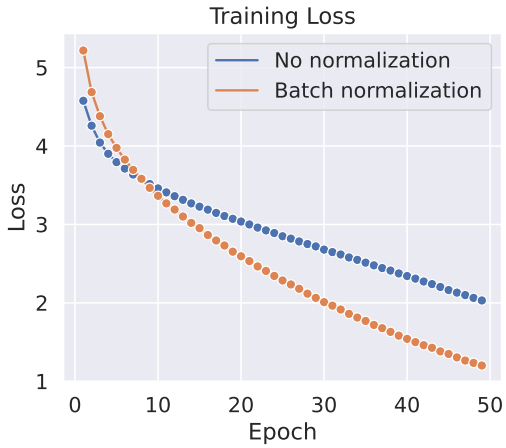

Figure 8. Training loss of MLP with (orange) and without (blue) batch normalization on the CIFAR100 dataset.

Rapidly mixing assumption. One limitation of our work is the rapidly mixing assumption that was used to establish the concentration of our results. While our experiments validated our results based on this assumption, it would be beneficial to prove that this assumption holds for a wide range of neural networks with batch normalization.

Training and optimization. While our focus of the current work was on random neural networks. In an elegant observation, Feng et al. (2022) demonstrate that the rank of input-output Jacobian of neural networks without normalization at initialization diminishes at an exponential rate with depth (Theorem 5), which implies changes in the input does not change the direction of outputs. In a remarkable observation, Yang et al. (2019) show the exact opposite for BN-MLP using a mean-field analysis (Theorem 3.10): any slight changes in the input lead to unbounded changes in the output. These results naturally raise the following question: Can we arrive at non-trivial results about input-output Jacobian at the infinite depth finite width regime?

The mean-field approach is also used to analyze the training mechanism. In particular, Chizat & Bach (2018) prove that gradient descent globally converges when optimizing single-layer neural networks in the limit of an infinite number of neurons. Although the global convergence does not hold for standard neural networks, insights from this mean-field analysis can be leveraged in understanding the training mechanism. For example, Daneshmand & Bach (2022) proves the global convergence of gradient descent holds for specific neural networks with a finite width, and two dimensional inputs in a realizable setting.

Exploring other normalizations. More research is needed for other normalization techniques, such as weight

normalization (Salimans & Kingma, 2016) or layer normalization (Ba et al., 2016) to understand the impact of these normalization techniques on the robustness and generalization of neural networks. Our findings highlight the power of mean-field theory for analyzing neual networks with normalization layers.

Extending to other architectures Our analyses are limited to MLPs. Extending our work to convolutional neural networks and transformers would enable us to analyze and enhance initialization for these neural networks. In particular, recent studies have shown that transformers suffer from the rank collapse issue when they grow in depth (Noci et al., 2022). A non-asymptotic mean-field theory may enable us to tackle this issue by providing a sound understanding of representation dynamics in transformers.

Overall, our results demonstrate that depth is not necessarily a curse for mean-field theory, but can even be a blessing when neural networks have batch normalization. The inductive bias provided by batch normalization controls the error propagation of mean-field approximations, enabling us to establish non-asymptotic concentration bounds for mean-field predictions. This result underlines the power of mean-field analyses in understanding the behavior of deep neural networks, thereby motivating the principle development of new initialization and optimization techniques for neural networks based on mean-field predictions.

Acknowledgments and Disclosure of Funding

Amir Joudaki is funded through Swiss National Science Foundation Project Grant #200550 to Andre Kahles. Also, we acknowledge support from the Swiss National Science Foundation (grant P2BSP3_195698), the European Research Council (grant SEQUOIA 724063), the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001(PRAIRIE 3IA Institute), and NSF TRIPODS program (award DMS-2022448).

References

Ba, J., Erdogdu, M., Suzuki, T., Wu, D., and Zhang, T. Generalization of two-layer neural networks: An asymptotic viewpoint. In *International Conference on Learning Representations*, 2019.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in Neural Information Processing Systems*, 2018.

Chizat, L. and Bach, F. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pp. 1305–1338, 2020.

Daneshmand, H. and Bach, F. Polynomial-time sparse deconvolution, 2022. URL https://arxiv.org/abs/2204.07879.

Daneshmand, H., Kohler, J., Bach, F., Hofmann, T., and Lucchi, A. Batch normalization provably avoids ranks collapse for randomly initialised deep networks. *Advances in Neural Information Processing Systems*, 33: 18387–18398, 2020.

Daneshmand, H., Joudaki, A., and Bach, F. Batch normalization orthogonalizes representations in deep random networks. *Advances in Neural Information Processing Systems*, 34, 2021.

de G. Matthews, A. G., Hron, J., Rowland, M., Turner, R. E., and Ghahramani, Z. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018.

Devroye, L., Mehrabian, A., and Reddad, T. The total variation distance between high-dimensional Gaussians. *arXiv preprint arXiv:1810.08693*, 2018.

Doeblin, W. Sur deux problèmes de M. Kolmogoroff concernant les chaînes dénombrables. *Bulletin de la Société Mathématique de France*, 66:210–220, 1938.

Dong, Y., Cordonnier, J.-B., and Loukas, A. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning*, pp. 2793–2803, 2021.

Eberle, A. Markov processes. *Lecture Notes at University of Bonn*, 2009.

Feng, R., Zheng, K., Huang, Y., Zhao, D., Jordan, M., and Zha, Z.-J. Rank diminishing in deep neural networks. *arXiv preprint arXiv:2206.06072*, 2022.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256, 13–15 May 2010.

Hanin, B. Correlation functions in random fully connected neural networks at finite width. *arXiv preprint arXiv:2204.01058*, 2022.

Hanin, B. and Nica, M. Finite depth and width corrections to the neural tangent kernel. *arXiv preprint arXiv:1909.05989*, 2019.

Hoffman, A. J. and Wielandt, H. W. The variation of the spectrum of a normal matrix. In *Selected Papers Of Alan J. Hoffman: With Commentary*, pp. 118–120. World Scientific, 2003.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pp. 448–456, 2015.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Jones, Galin L. and Hobert, James P. Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statistical Science*, pp. 312–334, 2001.

Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S. Self-normalizing neural networks. *Advances in Neural Information Processing Systems*, 30, 2017.

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in Neural Information Processing Systems*, 32, 2019.

Li, M., Nica, M., and Roy, D. The future is log-Gaussian: Resnets and their infinite-depth-and-width limit at initialization. *Advances in Neural Information Processing Systems*, 34:7852–7864, 2021.

Li, M. B., Nica, M., and Roy, D. M. The neural covariance sde: Shaped infinite depth-and-width networks at initialization. *arXiv preprint arXiv:2206.02768*, 2022.

Neal, R. M. *Bayesian Learning for Neural Networks*, volume 118. Springer Science & Business Media, 1995.

Noci, L., Anagnostidis, S., Biggio, L., Orvieto, A., Singh, S. P., and Lucchi, A. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. *arXiv preprint arXiv:2206.03126*, 2022.

Pastur, L. and Martchenko, V. The distribution of eigenvalues in certain sets of random matrices. *Math. USSR-Sbornik*, 1(4):457–483, 1967.

Pennington, J. and Worah, P. Nonlinear random matrix theory for deep learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Pennington, J., Schoenholz, S., and Ganguli, S. The emergence of spectral universality in deep networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 1924–1932, 2018.

Ramachandran, P., Zoph, B., and Le, Q. V. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

Rosenthal, J. S. Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 90(430):558–566, 1995.

Salimans, T. and Kingma, D. P. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in Neural Information Processing Systems*, 29, 2016.

Saxe, A. M., McClelland, J. L., and Ganguli, S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Xiao, L., Bahri, Y., Sohl-Dickstein, J., Schoenholz, S., and Pennington, J. Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks. In *International Conference on Machine Learning*, pp. 5393–5402, 2018.

Yang, G., Pennington, J., Rao, V., Sohl-Dickstein, J., and Schoenholz, S. S. A mean field theory of batch normalization. In *International Conference on Learning Representations*, 2019.

A. Proof of main theorems

A.1. A concentration bound for the empirical covariance matrix

The following analysis pertains to the deviation between the sample covariance matrix, normalized by the true covariance, and the identity matrix. For a collection of d independent identically distributed i.i.d. samples in $\mathbb{R}^d$, represented as $x_1, x_2, ..., x_d \in \mathbb{R}^n$, the sample covariance matrix C_d is given by:

$$C_d = \frac{1}{d}\sum_{i=1}^{d} x_i x_i^T. \tag{12}$$

The true covariance matrix C is defined as the expected outer product of the samples, or:

$$C = \mathbb{E}[x_i x_i^T]. \tag{13}$$

We are interested in bounding the deviation of C_d from the covariance matrix C in terms of their Frobenius norm (denoted as $\|.\|_F$), as outlined in the lemma below. Note that if activation σ is uniformly bounded $\sigma(x)^2 \leq Bx^2$, and ϕ is the batch-norm operator, then $\|\sigma(\phi(x))^2\| \leq B\|\phi(x)\|^2 \leq Bn$. Thus, activations applied after normalization layers obey the condition of Lemma 5. With this point in mind, we will prove the concentration result for vectors that are uniformly bounded by the same quantity.

Lemma 5. *Let $x_1, \ldots, x_d \in \mathbb{R}^n$ be i.i.d. random vectors with covariance $\mathbb{E}x_i x_i^\top = C$ and sample covariance $C_d := \frac{1}{d}\sum_{i=1}^{d} x_i x_i^\top$. If the vector norms are universally bounded such that $\|x_i\|^2 \leq nB$ holds almost surely, then for $t \lesssim \sqrt{d}$, the following is true:*

$$P\left(\|C_d - C\|_F \gtrsim t\varepsilon\right) \leq \exp(-t^2), \quad \varepsilon := \frac{Bn}{\sqrt{d}}. \tag{14}$$

Here, the probability is taken over the random vectors $x_1, \ldots, x_d$.

We will use the last lemma to prove Theorem 1.

Lemma 6. *Under the same conditions as Lemma 5, if the covariance matrix C is not degenerate, i.e., it does not possess zero eigenvalues, for $t \lesssim \sqrt{d}$ it holds*

$$P\left(\|C^{-1}C_d - I_n\|_F \gtrsim t\varepsilon\right) \leq \exp(-t^2), \quad \varepsilon := \frac{B\|C^{-1}\|n}{\sqrt{d}}. \tag{15}$$

Proof of Lemma 5. Recall that Bernstein's inequality provides an upper bound on the probability that the sum exceeds a certain threshold t. Given i.i.d. variables $X_1, \ldots, X_d$, it states that are uniformly bounded $|X_i| \leq B$ for all i, we have

$$\mathbb{P}\left(\frac{1}{d}\sum_{i=1}^{d} X_i \geq t\right) \leq 2\exp\left(-\frac{dt^2/2}{K^2 + Kt/3}\right), \tag{16}$$

where $t > 0$ and σ^2 is the variance of $\sum_{i=1}^{d} E[X_i^2] \leq dB^2$. Define $X_i := \|x_i x_i^\top - C\|_F^2$. We have

$$\begin{aligned}
\|x_i x_i^\top - C\|_F^2 &\leq \|x_i x_i\|_F + \|C\|_F && (17)\\
&\leq Bn + \|E x_i x_i^\top\|_F && (18)\\
&\leq Bn + E\|x_i x_i\|_F && (19)\\
&\leq 2Bn. && (20)
\end{aligned}$$

Thus, we can plug $K := 2Bn$ into the Bernstein's inequality to get

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{d}\sum_{i=1}^{d}\|x_i x_i^\top - C\|_F \geq t\right) \leq && (21)\\
&\exp\left(-\frac{dt^2/2}{4n^2B^2 + 2Bnt/3}\right). && (22)
\end{aligned}$$

Since $\|\cdot\|_F$ is convex, Jensen's inequality which implies that moving the averaging inside can only decrease its value, which in turn implies

$$\begin{aligned}
&\mathbb{P}\left(\|\frac{1}{d}\sum_{i=1}^{d} x_i x_i^\top - C\|_F \geq t\right) && (23)\\
&\leq \exp\left(-\frac{dt^2}{8n^2B^2(1 + \frac{t}{6nB})}\right). && (24)
\end{aligned}$$

We can now rename $t\sqrt{d}/(\sqrt{8}Bn)$ as t and use definition of sample covariance to conclude

$$\mathbb{P}\left(\|C_d - C\|_F \geq t\frac{\sqrt{8}Bn}{\sqrt{d}}\right) \leq \exp\left(-\frac{t^2}{(1 + \frac{t}{3\sqrt{2d}})}\right), \tag{25}$$

which can be restated as

$$P\left(\|C_d - C\|_F \gtrsim t\varepsilon\right) \leq \exp(-t^2), \quad \varepsilon := \frac{Bn}{\sqrt{d}}, \tag{26}$$

which holds if $t \lesssim \sqrt{d}$. □

Proof of Lemma 6. Consider transformed vectors $z_i := C^{-1/2}x_i$. Note that we have $\mathbb{E}z_i z_i^\top = C^{-1}C = I_n$. Thus, we can apply Lemma 5 on deviations of sample covariance of z_i's from I_n. Furthermore, we have $\|z_i\|^2 \leq \|C^{-1}\|\|x_i\|^2 \leq \|C^{-1}\|Bn$. So we can invoke Lemma 5 by setting B to $\|C^{-1}\|B$. □

A.2. Analyzing Gram Dynamics Around Fixed Points

Equipped with the results established so far, we now turn our attention to the dynamics of Gram matrices in relation to the total variation of the Multi-Layer Perceptron (MLP) Markov chain. In particular, we demonstrate a specific construction based on fixed-point G_*, and show that after one layer update the total variation distance is bounded.

Lemma 7 (Restated Lemma 2)**.** *Let $\hat{h} \in \mathbb{R}^{d\times n}$ be constructed by drawing its rows i.i.d. from $\mathcal{N}(0, G_*)$. Let $\hat{\mu}$ denote the distribution of $\hat{h}$. Given that fixed-point Gram matrix G_* is non-degenerate and the activation is uniformly bounded $\sigma(x)^2 \leq Bx^2$, then*

$$\|\hat{\mu} - T(\hat{\mu})\|_{tv} \lesssim \varepsilon^2 \ln(1/\varepsilon), \quad \varepsilon := \frac{n\|G_*^{-1}\|B}{\sqrt{d}}, \tag{27}$$

holds if B and $\|G_*^{-1}\|$ are non-zero.

Under the assumption of geometric contraction, irrespective of the initial distribution, the total variation distance to the stationary distribution contracts by $1-\alpha$, for some $\alpha > 0$, after one transition T. This result, together with Lemma 2, provides a tool to approximate the stationary distribution by a matrix constructed from the fixed-point Gram matrix G_*.

Lemma 8. *Let $\hat{\mu}$ denote the distribution of a random matrix in $\mathbb{R}^{d\times n}$, whose rows are drawn i.i.d. from $\mathcal{N}(0, G_*)$. Assuming the rapid mixing condition 1 holds with constant $\alpha > 0$, then*

$$\|\hat{\mu} - \mu_*\|_{tv} \lesssim \alpha^{-1}\varepsilon^2 \ln(1/\varepsilon), \quad \varepsilon := \frac{nB\|G_*^{-1}\|}{\sqrt{d}}. \tag{28}$$

We can finally the results about total variation into the context of Gram matrix dynamics through depth.

Lemma 9. *Let μ_ℓ denote the hidden representation of a BN-MLP, and $\hat{\mu}$ denote the distribution of a matrix whose rows are drawn from $\mathcal{N}(0, G_*)$. If hidden representations obey the rapidly mixing assumption with rate $1-\alpha$, for $\alpha > 0$, then*

$$\|\mu_\ell - \hat{\mu}\|_{tv} \lesssim (1-\alpha)^\ell + \alpha^{-1}\varepsilon^2 \ln(1/\varepsilon), \quad \varepsilon := \frac{nB\|G_*^{-1}\|}{\sqrt{d}}. \tag{29}$$

With the necessary lemmas in place, we are now ready to present our main theorem.

Theorem 10 (Restated Theorem 1)**.** *For an MLP chain G_ℓ that originates from a non-degenerate input G_0, and that has a non-degenerate fixed point $G_,$ and that obeys the rapidly mixing assumption with $\alpha > 0$, we have the following:*

$$\mathbb{P}(\|G_* - G\|_F \geq t) \lesssim t^{-2}(\|G_*\|^2(1-\alpha)^\ell + \alpha^{-1}\varepsilon^2 \ln(1/\varepsilon)), \tag{30}$$

with $\varepsilon := nB\kappa(G_*)/\sqrt{d}$.

The proof of the theorem relies on the following lemma bounds Gram matrix deviations by total variation.

Lemma 11. *Conditioned on Gram matrices $G_*, G \in \mathbb{R}^{n\times n}$, construct $h, \hat{h} \in \mathbb{R}^{d\times n}$ by drawing their rows i.i.d. from $\mathcal{N}(0, G)$ and $\mathcal{N}(0, G_*)$. If G_* is non-degenerate, the following hols for total variation between h and $\hat{h}$:*

$$tv(h, \hat{h}) \geq \frac{t}{100}\mathbb{P}(\|G_*^{-1}G - I_n\|_F^2 \geq t), \tag{31}$$

where the probability is defined over G.

The proof of this theorem follows directly from the lemmas we have established.

Proof of Theorem 10. We apply the total variation bound established in Lemma 9 and combine this with the lower bound stated in Lemma 11

$$\frac{t}{100}\mathbb{P}(\|G_*^{-1}G - I_n\|_F^2 \geq t) \tag{32}$$
$$\leq tv(h_\ell, \hat{h}) \lesssim (1-\alpha)^\ell + \alpha^{-1}\varepsilon^2 \ln(1/\varepsilon). \tag{33}$$

Omitting constants we have

$$\mathbb{P}(\|G_*^{-1}G - I_n\|_F \geq \sqrt{t}) \tag{34}$$
$$\lesssim t^{-1}((1-\alpha)^\ell + \alpha^{-1}\varepsilon^2 \ln(1/\varepsilon)). \tag{35}$$

By a change of variables we get

$$\mathbb{P}\left(\|G_*^{-1}G - I_n\|_F \geq t\right) \tag{36}$$
$$\lesssim t^{-2}((1-\alpha)^\ell + \alpha^{-1}\varepsilon^2 \ln(1/\varepsilon)). \tag{37}$$

Note that $\|G^* - G\|_F = \|G^*(G_*^{-1}G - I_n)\|_F$, which is bounded by $\|G_*\|\|G_*^{-1}G - I_n\|_F$. Thus

$$\mathbb{P}\left(\|G_* - G\|_F \geq t\right) \tag{38}$$
$$\leq \mathbb{P}\left(\|G_*^{-1}G - I_n\|_F \geq t/\|G_*\|\right) \tag{39}$$
$$\leq t^{-2}\|G_*\|^2((1-\alpha)^\ell + \alpha^{-1}\varepsilon^2 \ln(1/\varepsilon)), \tag{40}$$

where the last equation is due to equation 33. The above inequality obtains

$$\varepsilon := \frac{Bn}{\sqrt{d}}\|G_*\|\|G_*^{-1}\| \tag{41}$$
$$\mathbb{P}(\|G_* - G\|_F \geq t) \lesssim \tag{42}$$
$$t^{-2}\left(\|G_*\|^2(1-\alpha)^\ell + \alpha^{-1}\varepsilon^2 \ln(1/\varepsilon))\right). \tag{43}$$

Recall that $\|G_*\|\|G_*^{-1}\|$ encodes the ratio of largest to smallest eigenvalue of G_*, which is its condition number $\kappa(G_*)$. This finalizes the proof. □

Proof of Lemma 9. First, note that geometric contraction assumption implies

$$\|\mu_\ell - \mu_*\|_{tv} \le (1-\alpha)\|\mu_{\ell-1} - \mu_*\|_{tv} \le (1-\alpha)^\ell, \quad (44)$$

which by numerical inequality $1 - x \le \exp(-x)$ can be bounded by $\exp(-\alpha\ell)$. We can invoke Lemma 8 and triangle inequality for total variation to conclude the proof. □

Proof of Lemma 8. Recall that the rapidly mixing assumption implies that $\|T(\hat{\mu}) - \mu_*\|_{tv} \le (1-\alpha)\|\hat{\mu} - \mu_*\|_{tv}$. Furthermore, invoking Lemma 7, we have

$$\|T(\hat{\mu}) - \hat{\mu}\|_{tv} \lesssim \varepsilon^2 \ln(1/\varepsilon), \quad (45)$$
$$\implies \|\hat{\mu} - \mu_*\|_{tv} \lesssim \alpha^{-1}\varepsilon^2 \ln(1/\varepsilon), \quad (46)$$

where the last line is implied by the triangle inequality for total variation. □

Proof of Lemma 7. Let us explicitly construct $T(\hat{\mu})$. Recall that $\hat{\mu}$ describes distribution of $\hat{h}$ whose rows are drawn i.i.d. from $\mathcal{N}(0, G_*)$. Define $h := W\sigma(\phi(\hat{h}))$, where W is a Gaussian with i.i.d. elements $\mathcal{N}(0, 1/d)$. Thus, by construction, distribution of h follows $T(\hat{\mu})$. Our main proof strategy of upper bounding total variation between $\hat{h}$ and h is to bound it conditioned on the proximity of G to G_*.

Bounding deviations $\|G_*^{-1}G - I\|_F$. Recall that based on the fixed-point property of G_* we have

$$\mathbb{E}_{w\sim N(0,G_*)}\sigma(\phi(w))^{\otimes 2} = G_*. \quad (47)$$

Define sampled Gram of activations $G := \frac{1}{d}\sigma(\phi(\hat{h}))^\top\sigma(\phi(\hat{h}))$, which is equal in expectation to $\mathbb{E}G = G_*$. By construction of batch norm operator which maps every row to $\sqrt{n}$-sphere, and the uniform bound $\sigma(x)^2 \le Bx^2$, we can conclude that rows of $\sigma(\phi(\hat{h}))$ are always bounded by

$$\|\phi(x)\| \le \sqrt{n} \qquad \forall x \in \mathbb{R}^n \quad (48)$$
$$\implies \|\sigma(\phi(x))\| \le \sqrt{Bn}, \qquad \forall x \in \mathbb{R}^n \quad (49)$$
$$\implies \|\mathrm{row}_k(\sigma(\phi(\hat{h})))\|^2 \le Bn, \qquad \forall k. \quad (50)$$

This allows us to invoke Lemma 6 to conclude:

$$\mathbb{P}(\|G_*^{-1}G - I_n\|_F \ge t\varepsilon) \le \exp(-t^2), \quad (51)$$

where $\varepsilon = B\|G_*^{-1}\|n/\sqrt{d}$.

Bounding total variation $tv(h, \hat{h})$. Define set of matrices $N_t := \{M \in \mathbb{R}^{n\times n} : \|{G_*}^{-1}M - I_n\|_F^2 \le \varepsilon^2 t^2\}$. Observe that conditioned on G, h is equal in distribution to a matrix rows are drawn i.i.d. from $\mathcal{N}(0, G)$. Thus, we can decompose the total variation based on depending on G belonging to neighborhood of G_* or not

$$tv(h, \hat{h}) \le \mathbb{P}\left\{\|{G_*}^{-1}G - I_n\|_F^2 \ge t^2\varepsilon^2\right\} \quad (52)$$
$$+ \sup_{G\in N_t} tv(\mathcal{N}(0, G), \mathcal{N}(0, G_*)) \quad (53)$$
$$\lesssim \mathbb{P}\left\{\|{G_*}^{-1}G - I_n\|_F \ge t\varepsilon\right\} + \frac{3}{2}t^2\varepsilon^2, \quad (54)$$

where in the last line we use the upper bound on total variation between two Gaussian matrices from (Devroye et al., 2018). Plugging our result for deviation of G and G_* we have

$$tv(h, \hat{h}) \le \frac{3}{2}t^2\varepsilon^2 + \exp(-t^2), \quad (55)$$

which holds for all $t \lesssim \sqrt{d}$. In particular, we can set $t^2 := \ln(2/3\varepsilon^2)$ which implies

$$tv(h, \hat{h}) \lesssim \frac{3\varepsilon^2}{2}(1 + \ln(2/2\varepsilon^2)), \quad (56)$$

which omitting constants can be restated as

$$tv(h, \hat{h}) \lesssim \varepsilon^2 \ln(1/\varepsilon^2). \quad (57)$$

To finish the proof, observe that condition $t \lesssim \sqrt{d}$ translates to $\ln(1/\varepsilon^2) = O(d)$ which in turn requires $\varepsilon \gtrsim \exp(-d/2)$. Plugging the definition of ε we have $nB\|G_*^{-1}\| \gtrsim \sqrt{d}\exp(-d/2)$. Since the right-hand side is $o(1)$, and $n \ge 1$, this condition will always hold if the boundedness and conditioning are non-zero $B, \|G_*^{-1}\| > 0$. □

Proof of Lemma 11. Define set of matrices $N_t := \{M \in \mathbb{R}^{n\times n} : \|{G_*}^{-1}M - I_n\|_F^2 \le t\}$. We have

$$tv(h, \hat{h}) = \int_G \mathbb{P}(G)tv(\mathcal{N}(0, G), \mathcal{N}(0, G_*)) \quad (58)$$
$$= \int_{G\in N_t} \mathbb{P}(G)tv(\mathcal{N}(0, G), \mathcal{N}(0, G_*)) \quad (59)$$
$$+ \int_{G\notin N_t} \mathbb{P}(G)tv(\mathcal{N}(0, G), \mathcal{N}(0, G_*)) \quad (60)$$
$$\ge \int_{G\notin N_t} \mathbb{P}(G)tv(\mathcal{N}(0, G), \mathcal{N}(0, G_*)) \quad (61)$$
$$\ge \mathbb{P}(G \notin N_t) \inf_{G\notin N_t} tv(\mathcal{N}(0, G), \mathcal{N}(0, G_*)) \quad (62)$$
$$\ge \frac{t}{100}\mathbb{P}\left(\|G_*^{-1}G - I_n\|_F^2 \ge t\right), \quad (63)$$

where in the last line we have used the lower bound for total variation of multivariate Gaussians from (Devroye et al., 2018). □

B. Empirical Spectral Distribution for various activations

We repeat the experiment ($n = 20, \ell = 10$) in Figure 5 for various activations and various widths $d = 100, 200, 500, 1000$, and observe similar results. The results validate tighter concentration bounds with d established in the main Theorem.

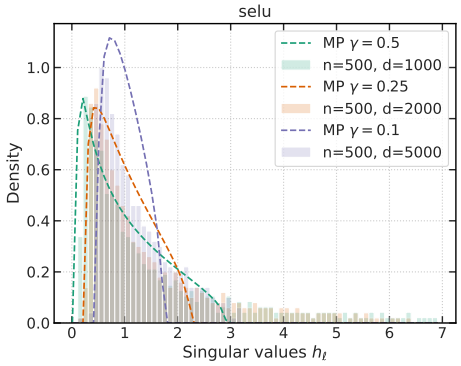

Figure 9. selu

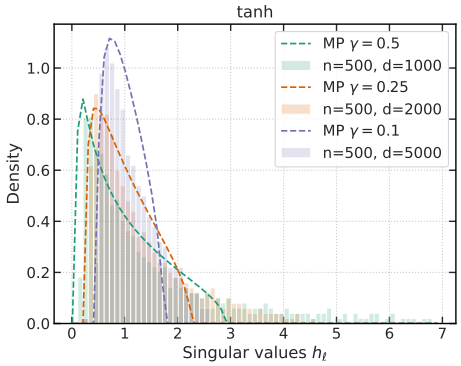

Figure 10. tanh

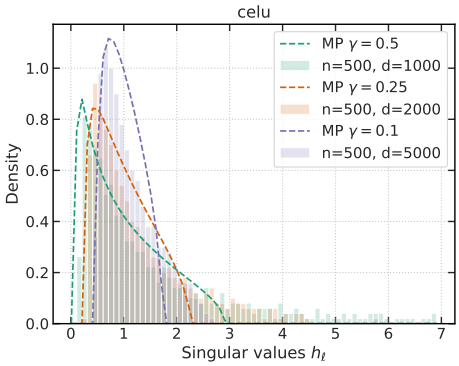

Figure 11. celu

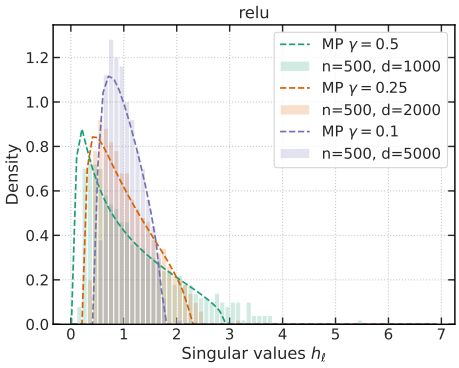

Figure 12. relu